Agents that Rationalize their Decisions

Luís Miguel Botelho

Department of Computer Science of ISCTE
Av. das Forças Armadas, Edificio ISCTE
1600 Lisboa, Portugal
luis@iscte.pt

Helder Coelho

Department of Computing Science
Faculty of Sciences
University of Lisbon
Bloco 5, Piso 1
Campo Grande
1700 Lisboa, Portugal
Helder.Coelho@di.fc.ul.pt

Abstract

This paper presents an introduction to a theory of rationalization for autonomous agents covering three kinds of rationalization: goal-based, belief-based, and assumption-based rationalization. We show that the COMINT model of decision making (Botelho & Coelho 1996) enables us to approach rationalization in a rather natural fashion. Namely, it presents an automatic (as opposed to deliberative and thoughtful) way of selecting the convenient sets of knowledge structures to rationalize the agent's decisions. This automatic mechanism relies on the concepts of activation and association, both of which are central to the model of memory (Botelho & Coelho 1995) underlying COMINT. We also give a formal definition of a rationalization basis and show how it is generated. Along the paper, we discuss some examples of personnel selection in organizations. Some aspects of the implementation of the model are also presented. Although rationalization serves a lot of purposes, this paper emphasizes its role in a multi-agent society.

1 Introduction

Human beings spend a great deal of time and effort making rationalizations of their choices and actions[1]. That is, during or after the process of arriving to a decision, we often build a rational justification for it. Sometimes we present those rational justifications to ourselves, sometimes we present them to an external audience. Rationalization serves a lot of purposes and it is used in a lot of situations (Pennington & Hastie 1988), (Tetlock 1992), (Shafir, Simonson & Tversky 1993), (Bobocel & Meyer 1994). However, the literature on Artificial Autonomous Agents has focused mainly on inter-agent communication (Cohen & Levesque 1995), distributed problem solving (Decker & Lesser 1995), reactive and situated behavior (Brooks 1991), and representation and manipulation of mental states (Konolige & Pollack 1993), but has devoted little effort to this important mental process.

We think rationalization is surely a useful topic for AI. From the cognitive science perspective, the study of rationalization enables us to understand an important mechanism used by human beings. From the perspective of building successful autonomous agents, rationalization provides a useful adaptive mechanism. It allows agents to get along with other agents (natural or artificial) to which they are accountable; and most important, it enhances the accessibility in memory of convenient solutions for given problems, improving the reactive and adaptive behavior of the agent. From the perspective of engineering friendly intelligent decision support systems, the very mechanism of rationalization may be used to provide the best way of conveying or explaining information to each user.

This paper aims at understanding some of the mechanisms underlying the rationalization process, how it can be implemented in computational agents and its role in multi-agent communities. Its fundamental contribution

[1]We use the word "rationalization" because it is an accurate and widely accepted term in Psychology. However, care must be taken not to confuse "rationalization of decisions" with "rational decisions". Often, a decision that needs to be rationalized is not a rational decision.

is a preliminary theory of rationalization for autonomous agents. The theory identifies three basic kinds of rationalization (goal-based, belief-based, and assumption-based), and describes common mental processes underlying them.

In agreement with current trends, we view rationalization as a globally deliberative (conscious) process in that it involves the adoption of a goal to rationalize a given decision, and resorts to explicit cognitive structures that specify the way to build it. However, it is conceivable that this globally deliberative process may have some automatic (not conscious) components.

We assume an agent may possess several distinct "knowledge islands" (i.e., sets of knowledge structures) that can be used to face the same problem. Consequently, we describe the rationalization process as a compound of three ingredients: (i) selection of a "knowledge island" to be used to generate the rationalization; (ii) generation of the rationalization using the knowledge structures contained in the "island" selected; and (iii) decision regarding the sufficiency and appropriateness of the rationalization generated. Although all these three components can be carried out in a deliberative fashion, we present alternative automatic approaches to sub processes (i) and (iii) drawing on properties of the COMINT model of decision making (Botelho & Coelho 1996, 1995). The deliberative nature of rationalization is preserved in sub process (ii). That is, the actual generation of a rationalization (once the "knowledge island" has been selected) is a deliberative process. We describe the way a rationalization may be generated (ii) and extend the original COMINT model to include this last capability.

In section 2, we present our approach to rationalization for autonomous agents in multi-agent environments; in section 3 we describe an example of rationalization; finally, in section 4 we present some final remarks.

2 What Is Rationalization

Rationalization is the process of building a rational justification for decisions, facts, actions or events, and presenting it to some audience. The audience to which the rational justification is presented may be the agent that rationalizes (internal audience) or another agent in a multi-agent environment (external audience). In what follows we use the term rationalization to refer both to the process and to the rational justification. In order to be easily accepted, the rationalization must suit the particular audience to which it is presented. Hence, the rationalization presented to a particular audience may be different from the rationalization presented by the same agent, for the same decision, but to a different audience. As a first example (Example 2.1), suppose a manager hires a particular secretary because of her salient physical attributes. If the manager has to present a rational justification of his decision to his supervisor, he'll probably say that she is very efficient and organized, and that she is capable of handling most problems without always having to ask what to do. However, if the manager wants to present a rationalization of his decision to the hired girl he might also refer to her being nice and attractive, both of those, very important qualities for the job.

Since a rationalization must suit a particular audience, it cannot be built during the decision-making process in the same way some systems keep a record of the reasons for each intermediate conclusion (Doyle 1981).

2.1 What Mental Processes Are Involved

The previous example makes it clear the rationalization process involves a reasoning mechanism through which an agent tries to find information in its memory or in the surrounding environment that enables it to support a given decision. Sometimes, if the agent doesn't have all the needed information, it may assume (even invent) some of it. For instance, when the manager tells his supervisor the girl is capable of handling most problems without always having to ask what to do, he probably doesn't know that for sure. In the rationalization presented, the manager just assumes or invents some hypotheses (assumption-based rationalization). This is fundamentally different from the process of generation of explanations traditionally used in expert systems and in intelligent tutoring systems. No one would expect such systems to create phony explanations for their conclusions or suggestions. This subsection addresses two important aspects of the rationalization process: the selection of the sets of knowledge structures ("knowledge islands") to be used to build the rationalization, and the problem of determining when to stop searching alternative rationalizations. Subsection 2.3 shows the way to build a rationalization from a given set of sentences (i.e., from the selected "knowledge islands").

2.1.1 Selection of Convenient Knowledge Structures

In Example 2.1, the manager presents a certain rationalization to his supervisor, and a different rationalization to the hired secretary. Both rationalizations presented and the actual decision were built from distinct sets of knowledge structures. This means decision makers may possess distinct sets of knowledge structures relevant to each problem. Therefore, the first step involved in the rationalization process is to choose among the possible sets of knowledge structures that support a given decision, one that is convenient for a given audience in a given context. There are two ways of approaching this problem. One is through a conscious deliberative process, the other is automatic and not conscious. The idea of a deliberative process implies the decision maker has something equivalent to a set of meta-rules specifying what rules to use if he or she was to rationalize a decision of a certain kind. Resorting to COMINT, we offer an alternative form of selection, one that is automatic and avoids all the complexities inherent to the explicit representation of rules and meta-rules (subsection 2.1.3). This approach also avoids a time consuming process of conscious selection of decision rules that most likely will force the decision maker to acknowledge having biased his or her former decision process in a self-serving way.

Of course, the model does not preclude a deliberative selection of an appropriate set of knowledge structures. We just present an alternative that seems plausible from a cognitive modeling point of view (it spares the agent the painful acknowledgment of having biased its information processing), besides being useful from the perspective of building intelligent agents (it is a more efficient and parsimonious approach).

It is worth noting that whether or not the selection of the set of knowledge structures is an automatic process, once it is selected, the reasoning performed thereafter is deliberative (conscious), in the sense that it involves the manipulation of explicit knowledge structures (e.g., rules and frames).

2.1.2 When to Stop?

After having arrived to a particular rationalization, what stops the decision maker to search another more convenient rationalization? Once more we may think of this in terms of a deliberative and conscious process aimed at evaluating the goodness of a particular rationalization, or in terms of an automatic process. Some have suggested that such decision should be thought about within a negotiation context. In this frame, the decision of when to stop depends on the perceived satisfaction of the audience. Although we agree with this suggestion it is important to emphasize three aspects. First, it doesn't forcefully imply a deliberative process is involved. Actually, someone may adapt his or her behavior to the perceived degree of satisfaction of the audience without consciously deliberating to do so. Second, the perceived degree of satisfaction of the audience may determine the agent's motivation to search information, therefore (indirectly) conditioning its motivation to try to find alternative rationalizations. Third, even if sometimes an agent uses a deliberative process to decide whether or not to stop trying alternative rationalizations, the same agent may do the same job automatically, in other situations. Therefore, we offer an automatic alternative approach to the decision of 'when to stop' based on the COMINT model of decision making (subsection 2.1.3). Once more, our proposal doesn't preclude the occurrence of deliberative processes.

2.1.3 Automatic Approaches

According to COMINT (Botelho & Coelho 1996, 1995), long term memory is an associative network represented by a directed labeled graph. Each node in the graph contains knowledge structures, and is characterized by an activation level. For our current purposes it suffices to say that more activated nodes are more accessible to the agent's information processing procedure.

When a problem (in particular, a rationalization problem) is put to the agent, its information processing procedure searches long term memory, in decreasing activation order, for a node that matches the problem -- the rationalization node, in case of a rationalization problem. If such a node is found, it gets activated and the nodes to which it is associated get activated too. The selection of convenient knowledge structures to build a rationalization depends on the rationalization node currently more activated and the nodes to which it is more strongly associated. There is nothing to think about: the agent just picks the rationalization node currently more activated (i.e., more accessible) in memory; if this node is not enough to produce the rationalization, other nodes are tried by activation order, until an answer for the current rationalization problem is found (or the agent runs out of motivation to search). If the motivation to search is still enough, the agent will try alternative rationalizations. Further more, if the agent is motivated to ignore undesired information, the information processing procedure may ignore some of the rationalizations found. In this way the COMINT model of decision making offers an automatic solution to both mentioned problems: (i) select the rationalization node more accessible, and (ii) stop when motivation has run out.

Due to the model of memory underlying the present work (SALT, (Botelho & Coelho 1995)) and to the conditions that trigger rationalization processes (e.g., fear of invalidity), it is likely that an agent improves its future performance after rationalization has taken place.

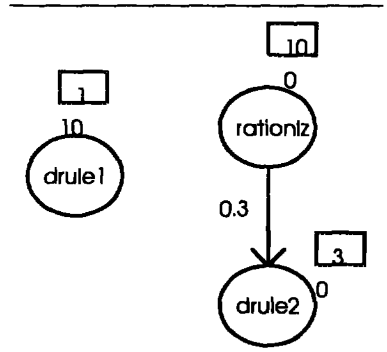

Figure 1 Dynamics of Long Term Memory

Suppose that, after a manager has made a poor decision due to a wrong decision rule (drule1, fig. 1), his supervisor asks him to justify his decision. Therefore the manager engages himself in a rationalization process trying to justify his former decision using an adequate decision rule, say drule2. After the manager has presented the rationalization to his supervisor, drule2 becomes more accessible in the manager's memory. Thereafter, if the manager faces a similar decision problem, he is more likely to use the adequate decision rule (drule2).

When the decision maker actually decides, drule1 (fig. 1) is used since it is the most activated of all (10 units). At the same time, the activation of both drule2 and rationlz is 0. When the decision maker is asked to rationalize his decision, some time has elapsed and the activation of drule1 has decayed to 1 (shown within a box above node drule1 in figure 1). In order to build the rationalization, the manager selects the node that matches the rationalization problem (node rationlz). As a result the activation of rationlz becomes 10 and the activation of drule2 becomes 3 (due to the association from rationlz to drule2 with strength 0.3). Therefore, if the agent faces the same kind of decision again, he will use decision rule drule2 instead of drule1. If drule2 is really a better decision rule for this kind of decision problem, the performance of the manager will improve. Notice that this improvement will only take place if the activation of the node (rationlz, in this case) is strongly associated to an adequate decision rule (drule2, in this case). In agreement with Philip Tetlock (Tetlock 1992) among others, it is likely that accurate or otherwise convenient decision rules are used when people have to rationalize their decisions. In such conditions, according to SALT, the rationalization node will become strongly associated to the accurate decision rule.

2.2 Three Basic Kinds of Rationalization

In this first attempt to build a theory of rationalization for autonomous agents, we introduce three kinds of rationalization: goal-based, belief-based, and assumption-based rationalization. In any of them, the agent tries to make the audience accept its decision. Hence, its relevance for multi-agent societies (e.g., in negotiation).

In the goal-based rationalization, the agent tries to show that its decision eases the achievement of one of the goals of the audience. For instance, it might say "Look, I've chosen Jack and let you choose Jane. That's what you've always wanted".

In the belief-based rationalization the agent tries to show that its decision can be derived from the beliefs of the audience. For instance, the agent might say "you know he is the best for the job".

It is important to stress that an agent may try to justify its decisions complying solely to the goals and beliefs of the audience and disregarding its own goals and beliefs. Therefore, it is perfectly possible that an agent produces a rationalization that is not consistent with its own goals and beliefs.

Finally, in the assumption-based rationalization the agent tries to show that its decision can be derived from assumed hypotheses. Hypotheses are assumed by abduction: they may be selected from the agent's own beliefs (as in "I think he is the most skilled"), or they may be totally new (e.g., when the decision maker is not aware of any fact that supports his or her decision and invents some phony reason). In this abductive process care must be taken not to assume an hypothesis inconsistent with salient beliefs of the audience. Of course, this constraint may hinder the generation of a rationalization, but we have to live with it. Assumption-based rationalization is fundamentally different from the process of generation of explanations in expert systems since it may assume inaccurate facts.

We have identified three basic categories of rationalization, but any actual rationalization process may combine some or all of them. Of course the success of the rationalization, (i.e., its degree of acceptance by the audience) depends on the accuracy of the beliefs the agent has about the audience. The construction of accurate models of the audience is not the concern of this paper, although we believe the very process of rationalization can be used to incrementally build models of other agents. If the audience accepts the rationalization presented, the agent may add the assumed hypotheses to the model of the audience.

In addition to the three kinds of rationalization presented, there are a lot of other possibilities. The agent may justify its decision invoking the existence of some rule or law making it choose a particular option (e.g., "by the rules, I must choose the most graduated"). We could call it norm-based rationalization. The agent may also say "my boss made me promote this guy". This could be termed power-based rationalization.

Besides the kinds of rationalization we assume to be generally available to all agents (e.g., goal-based, belief-based and assumption-based rationalization), an agent may develop specialized rationalization methods for concrete decisions and concrete audiences, instead of building a new rationalization each time it makes such decisions.

The remaining of this paper focuses only on goal-based, belief-based and assumption-based rationalization since they constitute powerful, general and representative mechanisms.

2.3 Definition of a Rationalization Basis

Through out the paper we have been talking about rationalization as a sentence or set of sentences presented by the decision maker to a given audience to justify his or her decision. In this subsection we give a formal definition of a rationalization basis. Informally, a rationalization basis is a set of reasons from which the agent is able of deriving its decision, and generating the actual rationalization. The rationalization is the sentences that are actually presented to the audience. The process by which an agent builds an actual rationalization from a rationalization basis is not the concern of this paper. Each reason in a rationalization basis contains an object sentence that is used to derive the decision. We call this object sentence, a decision support sentence. For instance, a rationalization presented to one's supervisor for the decision "hire Jessica Rabbit" may be something like "she is very efficient", a rationalization basis may be the set {"The supervisor believes someone should be hired if he or she is very efficient for the job", "One has assumed Jessica Rabbit is very efficient for the job"}, and the set of decision support sentences is {"Someone should be hired if he or she is very efficient for the job", "Jessica Rabbit is very efficient for the job"}.

2.3.1 Rationalization Basis

In subsection 2.1 we have described the mental processes involved in selecting the set of long term memory nodes used to face a rationalization problem. In what follows, we forget both the way those nodes were selected and the decision of whether or not to stop trying alternative rationalizations. We just consider the definition and the generation of a rationalization basis from the knowledge structures contained in the selected nodes. The organization and manipulation of memory (e.g., nodes, associations, activation and access methods) won't concern us here.

The formal definition of a rationalization basis for the decision δ presented by the decision maker γ to the audience α, involves the following concepts.

- A set of nodes selected by the agent to produce the rationalization (according to the COMINT model): $\{\eta_1, ..., \eta_r\}$.

- Given any node η (which is a complex data structure (Botelho & Coelho 1995)) we need a function kn to return the set of knowledge structures contained in it. $\Delta = kn(\eta_1) \cup ... \cup kn(\eta_r)$ is the set of knowledge structures contained in the nodes selected.

- A function for returning the decision support sentence contained in a given reason: ds. We assume the knowledge contained in the nodes of the agent's long term memory is represented using the language L. For the sake of simplicity we restrict L to be an extension of the first order predicate calculus with the usual modal operators Goal and Bel for goals and beliefs. In the present paper, it is not important to discuss the details and properties of L, nor to compare it to other languages of goals and beliefs. What really matters is the way a rationalization basis is defined and generated. Reasons are represented in the language J. If ψ and $\xi \in$ L and α is the audience of the rationalization then (i) $\psi \in$ J; (ii) Assumed(ψ) $\in$ J; (iii) Facilitates(Goal(α, ξ), ψ) $\in$ J; and (iv) nothing else belongs to J. Assumed(ψ) means the sentence ψ has been assumed (by the decision maker), and Facilitates(Goal(α, ξ), ψ) means ξ is one of the goals of the audience, and ψ facilitates its achievement. Given the languages L and J, the function ds from J to L is defined as follows:

ds(Bel(α, ϕ)) = ϕ
ds(Assumed(ϕ)) = ϕ
ds(Facilitates(Goal(α, ξ), ϕ)) = ϕ
ds(ϕ) = ϕ, otherwise

To clarify the definition of ds, notice for instance, ds(Bel(α, ϕ)) = ϕ means that if α believing ϕ is a reason for a given decision δ, then ϕ is a decision support sentence for δ. ds maps each reason of a rationalization basis into a formula of the knowledge representation language of the agent that may be used to derive the decision δ, that is, a decision support sentence for δ.

Given the above definitions, the set RB(δ)=$\{\psi_1, ..., \psi_n\}$ is a rationalization basis for the decision δ presented by the decision maker γ, to the audience α, iff:

(1) $\psi_i \in$ J (for all i=1, ..., n);

(2) $\{ds(\psi_1), ..., ds(\psi_n)\}$ is not known to be inconsistent by γ;

(3) $\{ds(\psi_1), ..., ds(\psi_n)\} \vdash_L \delta$; and

(4) RB(δ) may be generated according to the following rules:

(a) RB($\varphi_1 \wedge \varphi_2$) = RB($\varphi_1$) $\cup$ RB(φ_2)

(b) <u>Belief-based rationalization</u>. RB(δ)={Bel(α, ψ)} if Bel(α, ψ)$\in\Delta$ and σ is the most general variable substitution such that $\psi\sigma=\delta$; or RB(δ)={Bel(α, $\psi\leftarrow\varphi$)} $\cup$RB(ϕ) if Bel(α, $\psi\leftarrow\varphi$)$\in\Delta$, σ is the most general variable substitution such that $\psi\sigma=\delta$ and $\phi=\varphi\sigma$. $\psi\sigma$ denotes the application of σ to ψ. Informally, $\psi\sigma$ is an instance of ψ.

(c) <u>Rationalization that assumes beliefs of the decision maker</u>. RB(δ)={ψ} if $\psi\in\Lambda$, σ is the most general variable substitution such that $\psi\sigma=\delta$ and ψ is compatible with α according to γ (definition of compatibility in 2.3.2); or RB(δ)={$\psi\leftarrow\varphi$}$\cup$RB(ϕ) if ($\psi\leftarrow\varphi$)$\in\Delta$, σ is the most general variable substitution such that $\psi\sigma=\delta$, $\phi=\varphi\sigma$ and ($\psi\leftarrow\varphi$) is compatible with α according to γ;

(d) <u>Rationalization that assumes new hypotheses</u>. RB(δ)= {Assumed(ψ)} if ψ is assumed by abduction, σ is the most general variable substitution such that $\psi\sigma=\delta$ and ψ is compatible with α according to γ; or RB(δ)= {Assumed($\psi\leftarrow\varphi$)}$\cup$RB(ϕ) if ($\psi\leftarrow\varphi$) is assumed by abduction, σ is the most general variable substitution such that $\psi\sigma=\delta$, $\phi=\varphi\sigma$ and ($\psi\leftarrow\varphi$) is compatible with α according to γ;

(e) <u>Goal-based rationalization</u>. RB(ψ) = {Facilitates(Goal(α, ξ), ψ)} if Goal(α, ξ) $\in \Lambda$, ψ is compatible with α according to γ, and ψ facilitates the achievement of ξ (definition of facilitation in 2.3.3).

It should be emphasized however, that each of the preceding rules is actually used only if the agent that builds the rationalization chosen to do so. This point will be made more clear in the example described in section 3.

2.3.2 Compatibility

We define $B(\alpha)$ as the set of all accessible beliefs ascribed to the audience by the decision maker. That is, $B(\alpha)$ represents what the decision maker thinks the audience believes, in the moment the rationalization is being generated. $\beta \in B(\alpha)$ iff $Bel(\alpha, \beta) \in \Lambda$. ψ is compatible with α according to γ iff $\{\psi\} \cup B(\alpha)$ is not known to be inconsistent, by γ.

2.3.3 Facilitating the Achievement of a Goal

Intuitively, it does not make sense to help α achieve a given state of affairs ξ i.e., to facilitate the achievement of ξ, if α believes ξ to be the case or if α already believes ξ to be impossible. Thus the first condition for ψ to facilitate the achievement of ξ, is that neither $B(\alpha) \vdash \xi$ nor $B(\alpha) \vdash \neg\xi$. If none of the previous conditions holds, we say ξ is still possible. Given this concept of possibility, (i) ξ facilitates ξ iff ξ is still possible; and (ii) ξ facilitates φ iff φ is still possible, $(\varphi \leftarrow \varphi_1 \wedge \ldots \wedge \varphi_n) \in B(\alpha)$ and ξ facilitates at least one φ_i, with i = 1, ..., n.

The above definition of a rationalization basis has two important properties. First, each element of the rationalization basis specifies if it is assumed by abduction, if it is a belief of the decision maker himself, if it is a belief of the audience, or if it facilitates the achievement of a goal of the audience. In this way, the rationalization basis has all the necessary information for a rationalization to be properly constructed. Second, there is a systematic relation between a rationalization basis and a decision, given by property (3), i.e., the set of the decision support sentences of a rationalization basis for a given decision implies that decision. Furthermore, the rationalization basis can be computationally generated following rules (4:a-c).

3 Rationalizing Personnel Decisions

In this section we describe a realistic rationalization of a personnel selection decision. In our scenario (Example 3.1), Mr. Smith runs a small local newspaper with two reporters: Bill and Django. Mr. Smith is uncle of Bill and he wants to give the boy a chance. Therefore he decided to assign his own nephew to an interview of an important representative of a local gypsy community. Mr. Smith has to present a rationalization of his decision to his own nephew, who is a very proud young boy.

We have built a COMINT* agent (i.e., an agent designed according to the COMINT* model) that plays the role of Mr. Smith. COMINT* extends the original COMINT (Botelho & Coelho 1996) model to include the theory of rationalization presented in section 2. COMINT* is implemented as a Prolog program that includes an interpreter that reads a specification of an agent's long term memory, also written in Prolog. Fig. 2 depicts the relevant cognitive structures of Mr. Smith. The agent's long term memory is specified in a declarative fashion and doesn't contain any aspect of the automatic information processing mechanism of COMINT*.

```
link(bill, drule2, 0.8).
link(rationlz, drule1, 0.8).

node: bill
goal(bill, goodReporter(bill)).
bel(bill, goodReporter(X):-
   mostSkilled(X, _)).
deservesChance(bill).
relative(bill).

node: rationlz
rationalize(assign(X,T), X, R):-
   justify(assign(X,T),X, R,
         [bel, goal, assumed]).

node: drule1
assign(X, T):-
    interested(X, T),
   mostSkilled(X, T).

node: drule2
assign(X, T):-
    trustworthy(X),
   deservesChance(X).
trustworthy(X):-
   relative(X).
```

Figure 2 - Part of the agent's long term memory

The expressions link(bill, drule2, 0.8) and link(rationlz, drule1, 0.8) specify an association strength of 0.8 from node <u>bill</u> to decision rule 2, and from node <u>rationlz</u> to decision rule 1. The expression rationalize(assign(X, T), X, R) means that R is a rationalization of the decision assign(X, T), presented to audience X. Finally, the expression justify(assign(X, T), X, R, [bel,goal,assumed]) means that R is a justification of the decision assign(X, T) based on beliefs and goals of X, and also on assumed hypotheses (tried in the specified order). The power of building such a justification is embedded in the information processing procedure of the COMINT* model. This allows the programmers to avoid the burden of writing a full program to perform rationalizations. All that is needed is the specification of the kinds of rationalization desired and the order in which they should be tried. For instance, if the rationalization were to be based solely on the beliefs and goals of the audience, but excluding assumed reasons, the node <u>rationlz</u> would contain the rule rationalize(assign(X, T), X, R) :- justify(assign(X, T), X, R, [bel, goal]).

In what follows, we omit some details regarding the activation of nodes in long term memory, since such details are only relevant to the COMINT model of decision making (Botelho & Coelho 1996, 1995), but not to the rationalization process.

Before the rationalization process, Mr. Smith has decided to assign Bill to the interview because Bill is trustworthy and deserves a chance (decision rule 2). Therefore, the nodes <u>drule2</u> and <u>bill</u> get highly activated because they were both selected in order to generate the decision (see (Botelho & Coelho 1995, 1996) for details about the decision making process). After a while and some activation decay, the agent is asked to rationalize its decision: rationalize(assign(bill, interview), bill, R). Then the information processing procedure of the agent searches its long term memory for a node that matches this query, and it selects the node rationlz. As this node gets activated, the activation spreads to node drule1 through the association between the two. Therefore, decision rule 1 gets also highly activated. We assume the time elapsed between the decision and the rationalization is such that the activation of decision rule 2 is now less than the activation of decision rule 1. In order to rationalize assign(bill, interview), the agent's information processing procedure tries to find a belief of the form bel(bill, assign(X, T) :- Body), such that X and T are or can be instanciated by bill and interview, respectively (rule 4 (b), subsection 2.3). Since it cannot find such a belief, it tries to find a goal(bill, G) such that assign(bill, interview) facilitates the achievement of G (rule 4 (c)). Since it cannot find such a goal, it tries to assume a proposition of the form assign(X, T) :- Body, such that X and T are or can be instanciated by bill and interview, respectively. As it searches long term memory it finds drule1 (the most accessible at this moment): assign(X, T):- interested(X,T), mostSkilled(X, T) (rule 4 (c)). At this point of the rationalization, the agent has to rationalize both interested(bill, interview) and mostSkilled(bill, interview) (rules 4 (c) and (a)). Repeating the same process as before, it has to assume that interested(bill, interview) is the case (rule 4 (d)); as for mostSkilled(bill, interview), the agent finds that Bill has a goal that may be achieved if he were to believe that mostSkilled(bill, interview) is the case (rule 4 (c)): goal(bill, goodReporter(bill)) and bel(bill, goodReporter(X):-mostSkilled(X, _)). The rationalization basis is the outcome of the process just explained:

```
[ (assign(X, T) :- interested(X, T), mostSkilled(X, T)),
  (assumed(interested(bill, interview))),
  (facilitates(goal(bill, goodReporter(bill)),
  mostSkilled(bill, interview))) ]
```

According to the definition of function ds (subsection 2.3.1) the list of decision support sentences contained in the previous rationalization basis is

```
[ (assign(X, T) :- interested(X, T), mostSkilled(X, T)),
  (interested(bill, interview)),
  (mostSkilled(bill, interview)) ]
```

which implies the decision assign(bill, interview). The final rationalization would be something like "I have assigned you (Bill) to the interview because I have assumed you were interested in it and because you're the most skilled for the job". Assuming the agent has an accurate model of Bill, we argue this rationalization would be very well accepted by him. First and most important, Bill would realize that being considered the most skilled for the interview would enable him to achieve his goal of being a good reporter. Second, because the general rule "assign someone that is interested and the most skilled", and the hypothesis "Bill is interested" are not inconsistent with his beliefs. It is also worth noting that the decision rule invoked in the rationalization is different from the decision rule used to actually decide, but the mechanism governing the selection of one of them to decide and the other to rationalize is automatic as opposed to deliberative, thoughtful and conscious. This constitutes an important advantage of our approach. Finally, as a result of the rationalization process, the more accurate decision rule drule1 is now more accessible in the agent's long term memory. This means the agent is more likely to use it next time it has to assign a reporter to some task, improving its performance.

4 Final Remarks

We have presented a preliminary attempt to define a theory of rationalization for artificial autonomous agents. The theory focuses on three basic kinds of rationalization, goal-based, belief-based, and assumption-based, and suggests other general and specialized kinds of rationalization. We have showed how the COMINT model of decision making provides an automatic mechanism for two important mental processes involved in rationalization: choosing the convenient sets of knowledge structures from memory to build the rationalization, and determining when to stop searching alternative rationalizations. Finally, we have presented a formal definition of a rationalization basis and a set of rules to generate it, and we have provided a systematic relation between a rationalization basis and a decision. This systematic relationship enables us to view rationalization as a special kind of automated reasoning for computer agents.

Through out the paper our discussion about rationalization has focused mainly on its role to justify previous decisions to another agent in a multi-agent society. However, as was stressed in (Pennington & Hastie 1988), (Tetlock 1992) and (Shafir, Simonson & Tversky 1993), the same process may also be used to generate decisions (reason-based decision making).

As some have suggested, rationalization is strongly associated with affect. Indeed, if the agent is unable of building a rational justification for its decisions, it will feel negative affect. Hence, rationalization may be thought of as a self-regulation strategy that avoids some sources of negative affect.

In section 1 we have said that rationalization could also enhance the accessibility of accurate (or otherwise convenient) methods for handling a given problem, improving the reactive and adaptive behavior of the agent.

Subsection 2.1.3. explains the way the knowledge structures involved in a rationalization get more accessible in the agent's memory improving the likelihood of them being selected. As was pointed out in (Tetlock 1992), the decision rules used in rationalizations are likely to be pretty accurate because the need for rationalization arises when the decision maker fears his or her decisions may have negative personal consequences. Therefore, if a problem is handled using a knowledge structure involved in previous rationalizations, it is likely that the solution is more accurate, or convenient. However, sometimes rationalization gives rise to poor performance or decisions (Bobocel & Meyer 1994). Unfortunately, the approach described in this paper doesn't suggest any way to avoid the undesired effects of rationalization, yet.

Aspects for future investigation include the way an agent may learn during the course of building a rationalization (if it feels the need to observe the environment to check if a particular reason is actually true), and to study if it is more acceptable to assume facts (that could be directly defeated by observation) or rules, in the assumption-based rationalization.

Acknowledgments

The authors are grateful to Filipe Santos, Graça Gaspar, Luis Antunes and Pedro Ramos for many useful discussions and suggestions. This work was partially supported by contract PSCH/OGE/1038/95.

References

Bobocel, D.R. and Meyer, J.P. 1994. Escalating commitment to a failing course of action: separating the roles of choice and justification. *Journal of Applied Psychology* 79:360-363.

Botelho, L.M. and Coelho, H. 1995. A schema-associative model of memory. In Proceedings of the Fourth Golden West International Conference on Intelligent Systems. p81-85. S. Francisco, Calif.: International Society for Computers and their Applications.

Botelho, L.M. and Coelho, H. 1996. Information processing, motivation and decision making. In Proceedings of the Fourth International Workshop on Artificial Intelligence in Economics and Management, Tel-Aviv, Israel. Forthcoming.

Brooks, R. 1991. Intelligence without representation. *Artificial Intelligence* 47:139-159

Cohen, P.R. and Levesque, H.J. 1995. Communicative actions for artificial agents. In Proceedings of the First International Conference on Multi-Agent Systems, 65-72. S. Francisco, Calif.: AAAI Press

Decker, K.S. and Lesser, V.R. 1995. Designing a family of coordination algorithms. In Proceedings of the First International Conference on Multi-Agent Systems, 73-80. S. Francisco, Calif.: AAAI Press

Doyle, J. 1981. Making difficult decisions. Technical Report STAN-CS-81-864, Dept. of Computer Science, Stanford Univ.

Konolige, K. and Pollack, M.E. 1993. A representationalist theory of intention. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'93), 390-395. International Joint Conferences on Artificial Intelligence, Inc.

Pennington, N. and Hastie, R. 1988. Explanation-based decision making: effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition* 14:521-533

Shafir, E., Simonson, I. and Tversky, A. 1993. Reason-based choice. *Cognition* 49:11-36

Tetlock, P.E. 1992. The inpact of accountability on judgment and choice: toward a social contingency model. *Advances in Experimental Social Psychology* 25:331-376